基于 KNE-BPNN 的电务设备故障预测 *

李晨光 [1]，乔　帅 [1]，杨晓杰 [1]，解伟凡 [2]，李川子 [1†]，李俊红 [1†]

(1. 河北师范大学 数学与信息科学学院，石家庄 050024; 2. 东南大学 信息科学与工程学院，南京 211189)

摘　要：针对铁路电务设备故障频发、运行效率低且无有效故障预测方法等现实问题，提出一种基于 K-均值—邻域近似条件熵与 BP 神经网络（KNE-BPNN）的电务设备故障预测模型。首先，采用基于 K-均值聚类的样例约简算法约简设备故障决策表中的冗余样例；其次，运用邻域近似条件熵属性约简方法对样例约简后故障决策表中的非必要属性进行约简；最后，使用经过样例和属性约简后的样本集训练 BP 神经网络并进行模型预测，直到模型输出结果满足预设条件为止。实验结果表明 KNE-BPNN 故障预测模型的预测精度和泛化性能均满足电务设备管理的实际需求。

关键词：BP 神经网络；邻域粗糙集；近似条件熵；属性约简；故障预测；K-均值聚类算法

中图分类号：TP183　　doi: 10.3969/j.issn.1001-3695.2018.03.0201

Fault prediction for communication andsignal equipment based on KNE-BPNN

Li Chenguang[1], Qiao Shuai[1], Yang Xiaojie[1], Xie Weifan[2], Li Chuanzi[1], Qiao Shuai[1†], Li Junhong[1†]

(1. *College of Mathematics & Information Science Hebei Normal University*, *Shijiazhuang* 050024, *China*; 2. *School of Information science & Engineering*, *Southeast University*, *Nanjing* 211189, *China*)

Abstract: Based on the practical problems, such as frequent failures, low operating efficiency and lacking of effective fault prediction methods for railway communication and signal (C&S) equipment, a fault prediction model for C&S equipment based on K-means–neighborhood approximate conditional entropy and BP neural network (KNE-BPNN) is proposed. Firstly, a sample reduction algorithm based on K-means clustering is used to reduce the redundant samples in the equipment failure decision table. Secondly, using neighborhood approximate conditional entropy attribute reduction theory to reduce the non-essential attributes in the fault decision table after sample reduction. Finally, the BP neural network is trained by using the reduced sample set, and the model is trained until its output meets the expected requirement. The experimental results show that the prediction precision and generalization performance of the KNE-BPNN fault prediction model can meet the actual requirements.

Key words: neural network; neighborhood rough set; approximate conditional entropy; reduction attributes; fault prediction; K-means method

0　引言

电务设备作为铁路运输系统的主要设备之一，在铁路运输安全体系中发挥着重要的保障作用错误!未找到引用源。。近年来，我国铁路建设逐步完善，铁路固定资产投资逐年增加，铁路科技也在不断创新。现如今，我国铁路建设已经步入高速发展的道路。与此同时，对电务设备的稳定性和可靠性也提出了更高的要求。由于电务设备内部结构复杂，极易发生故障。为了减少故障的发生，确保铁路系统安全稳定地运行，研究在设备运行过程中的设备故障预测方法，预测设备故障的发生，从而及时对设备进行维修和维护，减少由于设备故障带来的人员伤害和经济损失是一个重要且现实的问题。

鉴于故障设备预测存在较高的研究价值，许多专家学者对这类问题进行了深入、广泛的研究，也产生了很多重要的研究成果。比如，利用支持向量机（SVM）、神经网络、灰色系统、模糊系统、时间序列、实时专家系统，等进行故障设备预测方面的研究错误!未找到引用源。。其中，人工神经网络具有自学习、自组织和非线性的能力，并且能够克服早期人工智能在模式识别和

收稿日期：2018-03-23；**修回日期：**2018-05-12　**基金项目：**国家自然科学基金资助项目（61573127，61502144）；河北省科技厅重点研发计划资助项目（16455702D）；河北师范大学基金资助项目（L2017B09，S2016Y13）

作者简介：李晨光（1993-），男，河北邯郸人，硕士研究生，主要研究方向为智能信息处理、机器学习（1281518477@qq.com）；乔帅（1994-），女，河北邢台人，硕士研究生，主要研究方向为机器学习、图像标注；杨晓杰（1990-），女，河北平山人，硕士研究生，主要研究方向为机器学习、图像识别；解伟凡（1997-），男，河北石家庄人，本科生，主要研究方向为深度学习；李川子（1995-），女，山西运城人，硕士研究生，主要研究方向为深度学习、图像处理；李俊红（1971-），女，山西运城人，副教授，博士，主要研究方向为模糊信息处理、深度学习.

非结构化处理等方面的不足，展示出其良好的智能特性错误!未找到引用源。。

在实际应用中，进行预测的样本数据往往会出现维度高，数据量大，关系复杂等不利于加工处理的情况。这样的数据输入神经网路预测模型容易造成模型收敛速度慢、拟合程度低、泛化能力不强等现象，而基于粗糙集理论的属性约简能够从复杂的数据中提取出潜在的、精度高的、分类性能好的数据错误!未找到引用源。。虽然粗糙集属性约简能够消除冗余数据，提高模型训练速度，但是在样本集中依然存在无效的、冗余的样例，使预测模型呈现过学习、过拟合等现象，而采用基于 K 均值聚类[5,6]的样例约简算法能够有效地约简冗余样例，加快模型的学习速度。

现阶段铁路电务部门对于电务设备的管理主要依赖人工方式进行，该方式存在管理工作繁琐，管理效率低下，不适合多用户共享等弊端。同时，铁路电务设备结构复杂，设备组成部件容易老化，极易造成设备故障。如何专业化地管理电务设备，提高电务设备的安全性和可靠性，已成为铁路电务部门急需解决的重要课题。

针对上述问题，给出基于样例和属性的设备故障数据约简方法，并在此基础上构建模型对电务设备故障进行预测。首先采用 K-均值聚类对决策表进行样例约简；其次，使用基于邻域近似条件熵错误!未找到引用源。对约简后的决策表进行属性约简；最后构建基于 KNE-BPNN（KNE-BPNN 即 K-均值和邻域近似决策熵的 BP 神经网络）的电务设备故障预测模型。实验结果表明，该模型在设备故障预测方面表现出良好的性能，可为后续电务设备智能管理提供有效思路和方法。

1　K-均值样例约简

真实的样本数据集往往存在很多无关样例，这类数据会使模型出现过学习、过拟合等情况，导致模型泛化能力降低。据此，采用基于 K-均值聚类的样例约简算法对样本集进行约简，消除样本集中的无关、冗余样例，提高训练后的模型的泛化能力。

1.1　K-均值聚类算法

K-均值聚类是一种传统的聚类算法，主要目标是将 n 个观测样本划分成 K 个类别，使得每个观测值与其所在聚类的均值之间的距离最小。该算法首先随机选择 K 个观测值作为初始聚类中心，并根据每个观测样本到各个聚类中心之间的距离将样本划分到距离最小的簇。然后计算每个簇的均值并将其更新成簇的聚类中心，迭代调整观测值的类别划分，直到各个簇的聚类中心不再发生改变。通常 K 均值聚类采用平方误差准则函数，该准则函数可以使 K 个聚类尽可能地聚集和收敛[5,6]。K-均值聚类算法流程如图 1 所示。

K-均值聚类算法的执行步骤如下：

输入：观测样本集 X 和聚类数目 K，其中 $X=\{x_1,x_2,...,x_n\}$。

输出：K 个聚类，每个聚类的平方误差最小。

a）令迭代次数 $i=1$，随机选择 K 个观测样本分别作为第 j 个类别的聚类中心 $C_j[i]$, $j=1,2,...,K$。

b)计算每个观测样本与所有聚类的类别中心之间的距离 $DIS(x_k,C_j[i]), k=1,2,...,n,$ 当 $DIS(x_k,c_j[i])=\min\{DIS(x_h,c_j[i]), h=1,2,...,n\}$, 则将 x_k 划分到第 j 个聚类。

c) 令 $i=i+1$, 更新聚类中心的值，$c_j[i]=\frac{1}{n_j}\sum_{k=1}^{n_j}x_k^{[j]}, j=1,2,...,K$，计算误差平方准则函数 J_e 的值 $J_e[i]=\sum_{j=1}^{K}\sum_{k=1}^{n_j}\|x_k^{[j]}-C_j[i]\|^2$。

d)如果 $|J_e[i+1]-J_e[i]|<\varepsilon$（$\varepsilon$ 为极小值，一般为10^{-4}），则算法终止，否则转第 b)步。

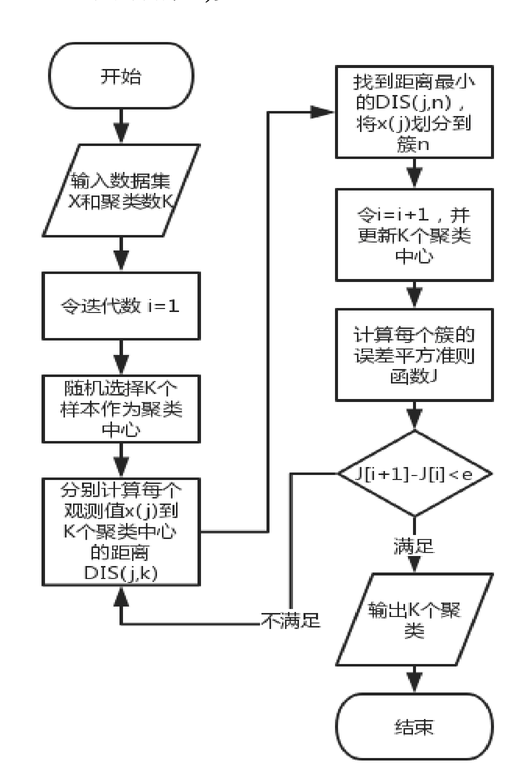


图 1　K-均值聚类算法流程图

1.2　样例约简算法

样例约简是基于 K-均值聚类算法进行约简的。首先根据样例集中的数据关系构建设备故障二维决策表 MDS（如表 1 所示，x_j 表示样例，$prop_i$ 为条件属性，D 代表决策属性，其中 0 表示代表未发生故障，1 表示发生故障)。其次，指定决策类别数为聚类数目 K，并根据决策类别划分数据子集(不包括决策属性) $MDS_k,(k=1,2,...,K)$，计算每个数据子集的均值，分别作为 K 个聚类的聚类中心。然后采用 K 均值聚类算法调整聚类和聚类中心，并输出 K 个聚簇。最后计算每个簇中的不同决策类别样例所占的比例，约简比例较少的样例。采用基于

K-均值聚类的样例约简算法能够有效地约简样本集中的无关、冗余样例，减少冗余样例对模型预测的影响，提升模型的泛化能力。

表 1 决策表

	$prop_1$	$prop_2$	$prop_3$	$prop_4$	D
x_1	1	2	1	1	0
x_2	2	3	0	2	1
x_3	3	3	1	1	1
x_4	2	4	1	2	0

基于 K-均值聚类的样例约简算法的具体实现如下：a)构建二维决策表，将决策表中的决策类别数设为聚类个数 K；b)按照决策类别划分数据子集且分别计算各数据子集的均值，并将均值指定为各个簇的聚类中心 C_k，其中数据子集不包括决策属性；c)按照 K-均值聚类算法调整各个聚类，最终得到 K 个簇；d)根据各个簇中不同决策类别所占的比例，约简比例较小的样本。样例约简的算法流程如图 2 所示。

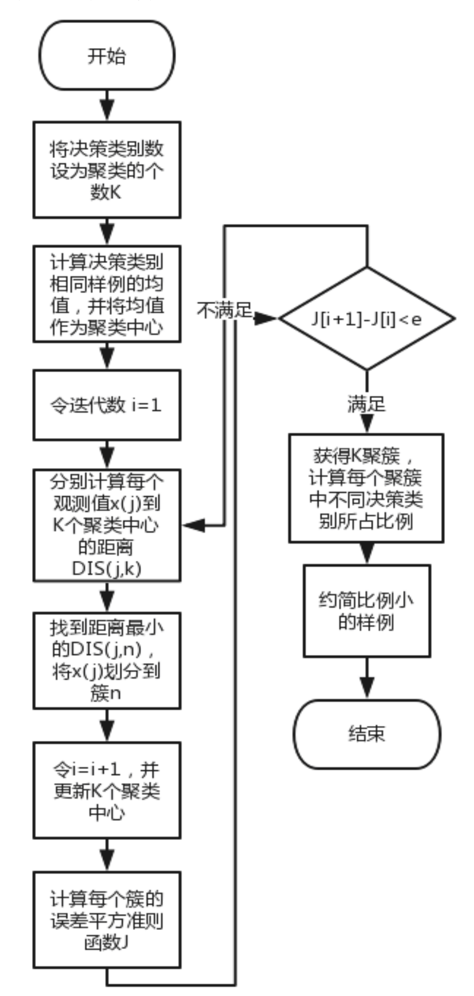


图 **错误!文档中没有指定样式的文字。**基于 K 均值聚类的样例约简算法流程图

2　基于邻域近似条件熵的粗糙集属性约简

2.1　粗糙集

粗糙集理论是波兰数学家 Z.Pawlak 于 1982 年提出的一种处理非确定性和不完整性知识的数学工具**错误!未找到引用源。**。采用规则提取方式进行属性约简是粗糙集理论的重要研究内容之一。现有的属性约简方法主要有基于保正域的决策粗糙集属性约简方法**错误!未找到引用源。**，基于差别矩阵的属性约简方法**错误!未找到引用源。**，以及信息观下的属性约简方法**错误!未找到引用源。**，等。但是，经典粗糙集理论仅适用于处理离散型数据，如果要对数值型数据进行约简，需要事先对其进行数据离散化操作，而离散化操作可能会导致重要信息的损失，影响后续约简操作的效果**错误!未找到引用源。**。邻域粗糙集约简模型建立在邻域等价关系基础上，不需要离散化操作就能有效处理数值型数据，最大程度地保持样本集的分类性能**错误!未找到引用源。**。

2.2　邻域粗糙集

用于分类学习的数据集可以定义为一个五元决策系统 $NDS=(U,C,D,V,f)$。其中 $U=\{u_1,u_2,...,u_n\}$ 是一个论域，为非空有限集，$C=\{c_1,c_2,...,c_n\}$ 表示对象的条件属性，为非空有限集，$D=\{d_1,d_2,...,d_n\}$ 表示对象的决策属性集合，$C\cap D=\phi$。$V=\bigcup_{a\in A}V_a$，其中 $A=C\cup D, V_a$ 是属性 a 的值域。$f:U\times(C\cup D)\to V$ 是一个信息函数，它为每个对象的每一个属性均设置一个信息值，即 $\forall a\in A, x\in U, f(x,a)\in V_a$。

定义 1[13]　对于任意一个对象 $u_i\in U, B\subseteq C, u_i$ 在属性集 B 上的邻域为

$$\delta_B(u_i)=\{u_j,u_j\in U,\Delta_B(\mathrm{u_i},\mathrm{u_j})\le\delta\},\delta\ge 0 \quad (1)$$

式中 Δ 表示两个对象之间的距离，通常用来衡量对象之间的相似度。

若 $A=C\cup D$,则在 N 维空间 $A=\{\mathrm{a_1,a_2,...,a_n}\}$ 中，样本 u_i 和 u_j 之间的 Minkowsky 距离为

$$\Delta_M(u_i,u_j)=[\sum_K^N|\mathrm{f(u_i,a_k)-f(u_j,a_k)}|^M]^{\frac{1}{M}} \quad (2)$$

其中 $f(u,a_k)$ 表示 u 在属性 a_k 上的取值。

常见的 Minkowsky 距离函数主要有三种，当 $M=1$ 时，此距离函数称为 Manhattan 距离；当 $M=2$ 时，称为 Euclidean 距离；当 $M=\infty$ 时，称为 Chebychev 距离，文中主要采用 Euclidean 距离。

定义 2[14,15]　在领域决策系统 $NDS=(U,C,D,V,f)$ 中，决策属性 D 决定的一个不可分辨关系为

$$IND(D)=\{(\mathrm{x,y})\subset\mathrm{U\times U}\mid\forall\mathrm{a\in S,f(x,a)=f(y,a)}\} \quad (3)$$

那么，$U/IND(D)=\{\mathrm{Y_1,Y_2,...,Y_m}\}$ 表示决策属性 D 对论域 U 的划分，$B\subseteq C$ 在论域 U 上的一个邻域关系为 N_B，$\delta_B(u)$ 为样本 u 在属性 B 上的邻域，则决策属性集 D 相对于 B 上的邻域上近似为

$$\overline{N_B}D=\bigcup_{i=1}^m\overline{N_B}Y_i \quad (4)$$

D 关于 B 的邻域下近似为

$$\underline{N_B}D=\bigcup_{i=1}^{m}\underline{N_B}Y_i \quad (5)$$

其中：$\overline{N_B}Y=\{u_i \mid \delta_B(u_i)\cap Y\neq\phi, u_i\in U\}$

$\underline{N_B}Y=\{u_i \mid \delta_B(u_i)\in Y, u_i\in U\}$

定义 3[7] 在邻域决策系统 $NDS=(U,C,D,V,f)$ 中，$\forall B\subseteq C, X\subseteq U$，$X$ 在 N_B 关系下的邻域近似精度为：

$$\gamma_B(X)=\frac{|\underline{N_B}X|}{|\overline{N_B}X|} \quad (6)$$

其中：$X\neq\phi$，$|X|$表示集合的势（或称基数，cardinal），$0\le Y_B(X)\le 1$。

2.3 基于邻域近似条件熵的粗糙集理论

定义 4[7] 对于一个邻域决策系统 $NDS=(U,C,D,V,f)$，$\forall B\subseteq C$，属性集 B 的邻域为 $\delta_B(u)$，决策属性 D 在论域 U 的划分为 $IND(D)=\{Y_1,Y_2,...,Y_m\}$，则属性 B 相对于决策属性 D 的邻域条件熵为

$$NCH=-\sum_{i=1}^{m}\sum_{j=1}^{n}[\frac{|\delta_B(u_j)\cap Y_i|}{|U|}\times\log_2\frac{\delta_B(u_j)\cap Y_i}{|\delta_B(u_j)|}] \quad (7)$$

定义 5[7] 对于一个邻域决策系统 $NDS=(U,C,D,V,f)$，由定义 1 可知 $\forall B\subseteq C$，属性集 B 的邻域为 $\delta_B(u_i)$，决策属性 D 在论域 U 的划分为 $IND(D)=\{Y_1,Y_2,...,Y_m\}$。由定义 3 可知 $\gamma_B(Y)$ 是 Y 在邻域条件关系 N_B 下的邻域近似精度，那么决策属性 D 相对于属性 B 的邻域近似条件熵为

$$NACH(D|B)=-\sum_{i=1}^{m}\{(1-\gamma_B(Y_i))\times \sum_{j=1}^{n}[\frac{|\delta_B(u_j)\cap Y_i|}{|U|}\times\log_2\frac{\delta_B(u_j)\cap Y_i}{|\delta_B(u_j)|}]\} \quad (8)$$

性质 1 错误!未找到引用源。 给定一个邻域决策系统 $NDS=(U,C,D,V,f)$，由定义 1 可知 $\forall B\subseteq C$，属性集 B 的邻域为 $\delta_B(u_i)$，决策属性 D 在论域 U 的划分为 $IND(D)=\{Y_1,Y_2,...,Y_m\}$，那么邻域近似条件熵满足 $0\le NACH(D|B)\le n\log_2 n,(n=|U|)$:

a）当且仅当 $\forall Y_j\in U/IND(D),|Y_j|=1,\forall u_i\in U,\delta_B(u_i)=U$ 时，

$$NACH(D|B)=MAX\{NACH\}=n\log_2 n$$

b）当 $\gamma_B(X)=\frac{|\underline{N_B}X|}{|\overline{N_B}X|}=1$，即所有对象都是正域中的元素时，$NACH(D|B)=MIN\{NACH\}=0$。

性质 2 错误!未找到引用源。 在邻域决策表系统 $NDS=(U,C,D,V,f)$ 中，若 $M,N\subseteq C$ 且 $M\subseteq N$，那么 $NACH(D|M)\ge NACH(D|N)$。

性质 3 错误!未找到引用源。 在邻域决策表系统 $NDS=(U,C,D,V,f)$ 中，若 $M\subseteq C$ 且 $a\in M$，当 $NACH(D|M)\ge NACH(D|M-\{a\})$ 时，认为属性 a 相对于 B 来说是不必要的。

定义 6 错误！未找到引用源。 在邻域决策表系统 $NDS=(U,C,D,V,f)$ 中，若 $M\subseteq C$，当 $NACH(D|M)=NACH(D|C)$，并且 $\forall m\in M, NACH(D|M-\{m\})>NACH(D|C)$ 时，称 M 是 C 相对于 D 的一个约简。

定义 7 错误！未找到引用源。 给定邻域决策系统 $NDS=(U,C,D,V,f)$，$\forall m\in C$，m 相对于 C 关于 D 的内部属性重要度为

$$IMP(m,C,D)=NACH(D|C)-\{m\})-NACH(D|C) \quad (9)$$

定义 8 错误!未找到引用源。 对于一个邻域决策系统 $NDS=(U,C,D,V,f)$。当 $\forall a\in C$，如果 $NACH(D|C)-\{a\})>NACH(D|C)$，即 $IMP(a,C,D)>0$，则称 a 是 C 相对于 D 的一个核属性。

定义 9 错误!未找到引用源。 给定邻域决策系统 $NDS=(U,C,D,V,f)$，若 $N\subseteq C$，当 $\forall n\in C-N$ 时，n 关于 D 的外部属性重要度为：

$$IMP(n,N,D)=NACH(D|N)-NACH(D|N\cup\{n\}) \quad (10)$$

2.4 基于邻域近似条件熵属性约简算法

通过邻域粗糙集相关理论获得所有条件属性集合的最小约简是一个 NP 问题，目前解决此类问题的一个有效方式是采用启发式的方法进行约简。启发式添加策略首先获取条件属性的核属性，然后循环比较各个核属性的属性重要性，将属性重要性 $IMP(a,C,D)$ 最大的元素添加到约简集 B 中，即 $B=B\cup\{a\}$，直到 $NACH(D|B)=NACH(D|C)$ 为止。下面是基于邻域近似条件熵的粗糙集属性约简算法的基本步骤：

输入：邻域决策系统 $NDS=(U,C,D,V,f)$ 和阈值 $\delta\ge 0$。

输出：邻域决策系统最小约简集 B。

a)初始化约简集 B，令 $B=\phi$。

b) $\forall u_i\in U$，采用 Euclidean 距离计算 u_i 在条件属性集 C 下的邻域 $\delta_C(u_i)$。

c)计算决策属性 D 关于 C 的邻域近似条件熵 $NACH(D|C)$。

d) $\forall a_i\in C$，计算 $IMP(a_i,C,D)$，若 $IMP(a_i,C,D)>0$，则将 a_i 添加到约简集 B 中，即 $B=B\cup\{a_i\}$。

e) 如果 $B\neq\phi$，计算 $NACH(D|B)$，若 $NACH(D|B)=NACH(D|C)$，则执行第 g)步，否则转到第 6 步执行。若 $B\neq\phi$，则直接执行第 f)步。

f) $\forall a_i\in C-B$，计算 a_i 的外部属性重要度 $IMP(a_i,B,D)$，若 a_k 满足 $IMP(a_k,B,D)=\max\{IMP(a_i,B,D),a_i\in C-B\}$（如果有多个属性满足该条件，则随机选择其中一个属性），则令 $B=B\cup\{a_k\}$，并跳转到第 e)步。

g)输出属性约简集 B，结束算法。

基于邻域近似条件熵的粗糙集属性约简算法流程如图 3 所示

3 基于 KNE-BPNN 的故障预测模型

基于 KNE-BPNN 故障预测模型的具体步骤如下：

a)利用设备历史故障监测数据构造初始故障决策表；

b)采用 K-均值聚类方法将决策表中的样例聚成两类（样本数据决策结果仅为发生故障和未发生故障两种），计算每个簇中不同决策类别样例的比例，约简比例较小的样例；

c)使用邻域近似条件熵约简算法对样例约简后的决策表进行属性约简，去除决策表的冗余属性；

d)经过 b)c)后产生一个新的约简决策表。并将其中的数据输入 BP 神经网络，经过反复训练和参数调优，确定神经网路每层的设置以及每层神经元节点的个数，最后利用调试好的神经网络对设备实施故障预测。

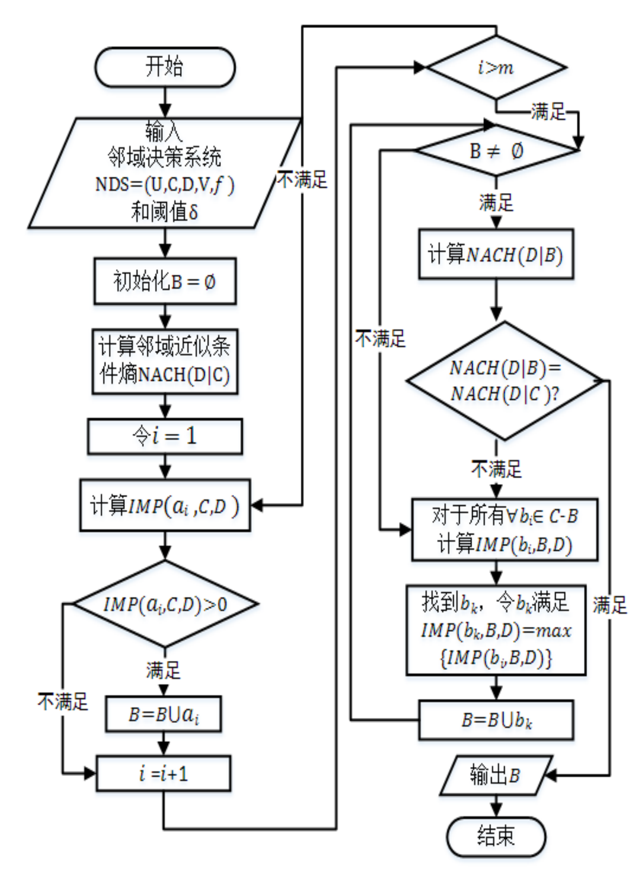


图 3　基于邻域近似条件熵的粗糙集属性约简算法流程图

基于 KNE-BPNN 的故障预测模型采用 K-均值样例约简和基于邻域近似条件熵属性约简算法对故障决策表进行约简操作，能够有效地约简冗余数据，确保样本数据保留更好的分类性能。而约简后的样本集输入神经网络后，能够提升神经网络的训练效率，避免模型出现过学习、过拟合状况。图 4 为基于 KNE-BPNN 故障预测模型工作流程，具体步骤如下：

a)故障数据决策表构建。将收集到的电务设备故障监测数据通过添加缺失值，替换一些异常值，最终初始化成一个二维的故障决策表。

b)样例约简。采用 K-均值聚类方法对初始故障决策表按照决策表中样例的决策类别数 N 将样例聚成 N 个簇，然后依次计算每个簇中不同决策类别的样例集所占的比例，并将比例较小的样本剔除，从而完成故障决策表的样例约简。

c)属性约简。首先计算故障决策表中所有条件属性集的邻域近似条件熵，初始化属性约简集 B 为空，并将所有内部重要度大于 0 的属性添加到约简集中。然后迭代计算属性约简集 B 的邻域近似条件熵，判断其是否与所有属性的近似条件熵相等。若相等，则输出；否则，计算除属性约简集以外所有条件属性的外部重要度，并将外部重要度最大的属性添加至属性约简集中，并再次进行迭代，直到属性约简集和条件属性集近似条件熵相等为止。

d)神经网络预测模型构建（图 5 为神经网络的结构图）。首先根据样本的维度和样本的类别数分别确定输入层节点数和输出层节点数。其次根据样本数据量和维度两个因素确定隐含层层数，然后为隐含层神经元设置节点数范围(例如设置隐含层的神经元个数 $N \in (50]$ 且 N 为正整数)，从该范围内均匀取数，设置为该隐含层的神经元节点数。通过反复实验确定最终的隐含层神经元个数。

e)输出结果分析。将样本测试集输入训练好的神经网络模型中，最终得到样本故障情况的预测结果。对输出结果进行分析，若达不到预测目标，则继续调整模型中的超参数或增加训练的次数，直到输出结果符合预设条件为止。

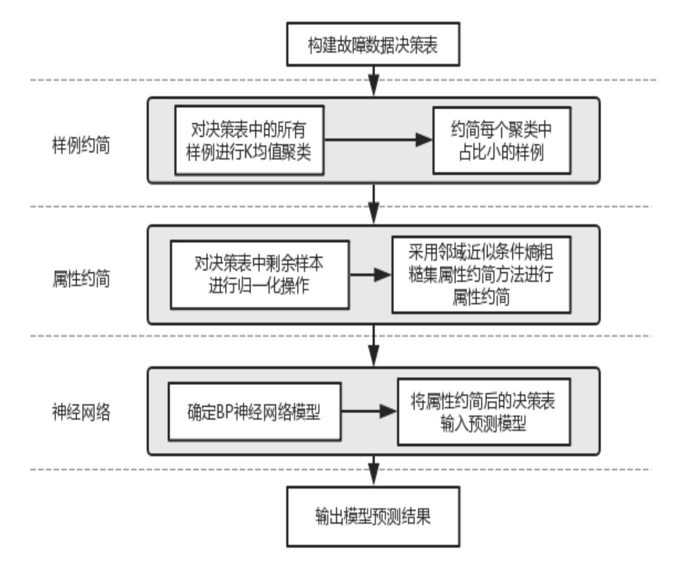


图 4　基于 KNE-BPNN 的故障预测模型

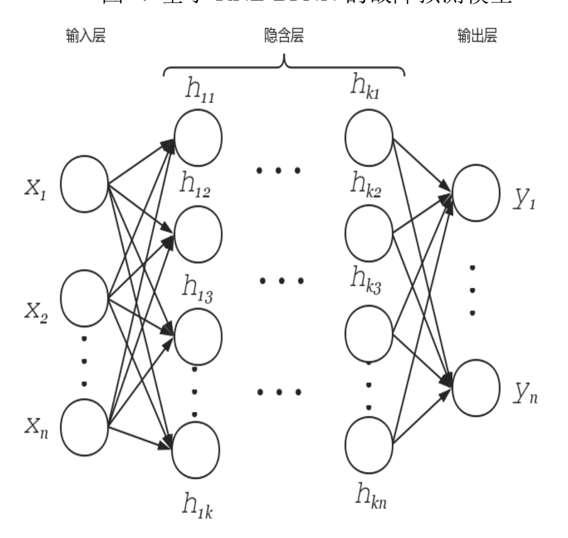


图 5　神经网络的结构图

4 基于 KNE-BPNN 故障预测实例分析

4.1 初始化决策表

文中使用铁路道岔设备电路的监测数据作为 KNE-BPNN 故障预测系统的样本集，如表 2 所示。将表 2 中的数据初始化成一个决策表 $NDS=(U,C,D,V,f)$，其中条件属性 C={温度，供电电压，发送电压，发送频率，接收 1.1 电压，接收 1.2 电压，干扰电压，故障情况}，决策属性 D={故障情况}（发生设备故障为 1，设备运转正常为 0）。并非所有初始决策表中的因素都会对道岔设备故障产生影响，而通过 K-均值样例约简以及邻域近似条件熵属性约简能够最大限度地保留决策系统中对道岔故障具有较大影响的样本。使用约简后的样本进行设备故障预测能够提高模型的预测精度和泛化能力。

表 2 铁路道岔设备故障决策表（部分数据）

编号	温度	供电电压	发送电压	发送频率	接收 1.1 电压	接收 1.2 电压	干扰电压	故障情况
1	28	201.5	58.67	9.85	5.07	8.55	0.63	1
2	25.5	205.23	55.72	9.75	8.92	6.95	0.76	1
3	25.8	192.92	59.94	9.66	8.76	8.16	0.72	1
4	29.9	225.37	56.14	9.69	5.01	8.91	0.59	0
5	29.7	221.47	56.45	9.58	4.89	4.6	0.47	1
6	27.2	212.87	58.16	9.56	8.54	5.07	0.68	1
7	29.4	217.7	56.19	9.81	6.33	7.06	0.87	1
8	26.3	208.33	50.86	9.58	5.94	8.27	0.86	0
9	25	214.34	57.99	9.31	8.22	8.5	0.5	1
10	24.1	207.77	50.42	9.45	5.96	8.62	0.78	0
11	28.6	238.51	51.7	9.79	6	8.34	0.74	0
12	20.5	190.59	56.48	9.11	4.88	8.42	0.48	1
13	29.5	216.3	59.09	9.08	6.57	6.33	0.71	0
14	23.7	216.8	54.19	9.16	8.76	6.45	0.66	0
15	20.9	227.7	57.83	9.55	7.53	6.1	0.6	0

4.2 K-均值样例约简

在采用 K-均值样例约简算法约简样例之前，应事先确定好聚类的个数 K 以及每个簇的聚类中心 C_k。由 K-均值样例约简算法可知，聚类个数 K 应设置为决策表决策属性的类别数。而为了使大多数样本能够聚类到样本本身所属的类别，先对样本集进行标准化处理，也就是对每个样本按照去均值和方差进行缩放操作，即

$$\hat{x}=\frac{x-x_mean}{x_std}$$

(11)

其中: x_mean 为均值， x_std 为标准差。

其次，将样例集按照决策类别划分子集，然后计算每个子集的均值，并将这些均值作为各个簇的聚类中心 C_k。即

$$\begin{cases} C_k=(c_{1k},c_{2k},\dots,c_{nk}) \\ c_{ik}=\frac{1}{M_k}\sum_{m=1}^{M_k}x_{mi}, x_m\in\{x\in U \mid x_d=D_k\} \end{cases} \quad (12)$$

其中 $k=0,1,\dots,K-1, K\ge 2$， n 为条件属性个数， M_k 为决策类别为 D_k 的样例集的个数，x_m 属于由决策类别为 D_k 的样例组成的样例集。

表 3 标准化故障决策表

编号	温度	供电电压	发送电压	发送频率	接收 1.1 电压	接收 1.2 电压	干扰电压	故障情况
1	-0.25	2.35	0.21	-0.52	-0.59	-0.54	-0.66	1
2	-0.28	2.37	0.16	-0.52	-0.53	-0.56	-0.65	1
3	-0.28	2.35	0.26	-0.54	-0.55	-0.56	-0.68	1
4	-0.24	2.38	0.11	-0.51	-0.58	-0.52	-0.63	0
5	-0.23	2.37	0.13	-0.50	-0.57	-0.57	-0.63	1
6	-0.27	2.37	0.17	-0.52	-0.53	-0.58	-0.64	1
7	-0.24	2.37	0.13	-0.51	-0.56	-0.55	-0.64	1
8	-0.26	2.38	0.10	-0.50	-0.56	-0.52	-0.63	0
9	-0.30	2.37	0.17	-0.52	-0.54	-0.53	-0.65	1
10	-0.29	2.39	0.10	-0.50	-0.55	-0.51	-0.63	0
11	-0.26	2.40	0.03	-0.50	-0.55	-0.52	-0.61	0
12	-0.33	2.35	0.24	-0.51	-0.58	-0.52	-0.65	1
13	-0.24	2.36	0.17	-0.53	-0.56	-0.56	-0.64	0
14	-0.31	2.38	0.12	-0.51	-0.51	-0.55	-0.63	0
15	-0.35	2.38	0.14	-0.50	-0.52	-0.54	-0.61	0

通过分析故障决策表的决策属性 D，可以明显观察到决策属性 D 有发生设备故障和设备运转正常两种情况，分别用 1 和 0 来表示，因此聚类个数 $K=2$；接着对样本集进行标准化处理（样本标准化处理结果如表 3 所示，样本数据均保留两位小数），数据标准化既能保证各个属性之间的差别依然存在，又平衡了各个属性之间的量级。然后按照决策类别划分子集，计算各个子集的均值，将其作为每个簇的聚类中心。计算可得

$C_0=(-0.28,2.38,0.11,-0.51,-0.55,-0.53,-0.63)$

$C_1=(-0.27,2.36,0.18,-0.52,-0.56,-0.55,-0.65)$。

确定聚类中心后，对样例进行聚类操作，直到聚类中心不再发生变化为止。由图 6 和表 4 可知，在 15 个样例中，样例编号为 5，7 和 13 的样本被错误划分，因此为了避免模型出现过学习，过拟合现象，应将这 3 个样例去除。

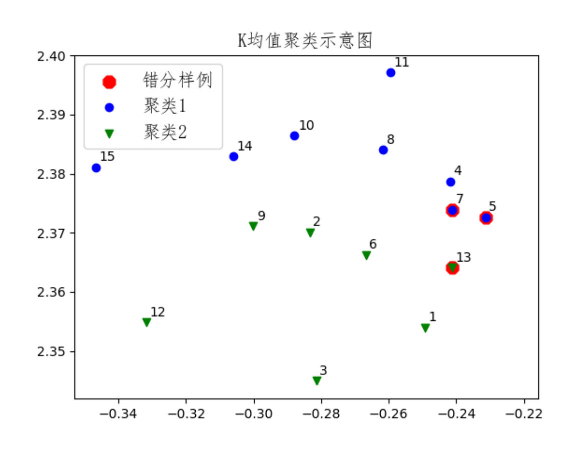


图 6　聚类结果示意图

样例的聚类结果如表 4 所示。

表 4 样例聚类分析结果

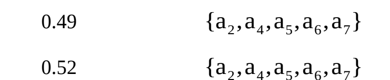
编号	实际类别	聚类类别	编号	实际类别	聚类类别
1	1	1	11	0	0
2	1	1	12	1	1
3	1	1	13	0	1
4	0	0	14	0	0
5	1	0	15	0	0
6	1	1			
7	1	0			
8	0	0			
9	1	1			
10	0	0			

4.3　邻域近似条件熵属性约简

样例约简操作完成之后，为防止模型训练速度过慢，学习率低的情况发生，采用邻域近似条件熵属性约简算法对条件属性进行约简。由表 3 可知在设备故障决策表有 7 个条件属性，在进行属性约简操作时，通过调整阈值δ能够控制约简属性的个数。一般情况下，约简后的属性子集个数应大于原始属性集的一半，而且属性子集应尽可能的小。所以属性集的个数$M_c \in [4,5]$，且M_c为整数。经计算满足要求的$\delta \in [0.38, 0.52]$，将$[0.38, 0.52]$以 0.03 为间隔分成 5 份，分别为$[0.38, 0.41)$，$[0.41, 0.44)$，$[0.44, 0.47)$，$[0.47, 0.50)$，$[0.50, 0.52]$。然后，从每个区间内取一个随机数，最终得到 5 个阈值，即$\delta_1 = 0.39$，$\delta_2 = 0.43$，$\delta_3 = 0.44$，$\delta_4 = 0.49$，$\delta_5 = 0.51$。最后，将各个阈值约简后的样例集，依次输入神经网络，并比较各个阈值所对应的预测精度，选择精度最高的属性约简子集作为最终的约简结果。

表 5 不同阈值的属性约简结果

阈值	属性约简集
0.39	$\{a_1, a_2, a_3, a_7\}$
0.43	$\{a_1, a_2, a_3, a_7\}$
0.44	$\{a_1, a_2, a_3, a_4\}$



0.49	$\{a_2, a_4, a_5, a_6, a_7\}$
0.52	$\{a_2, a_4, a_5, a_6, a_7\}$

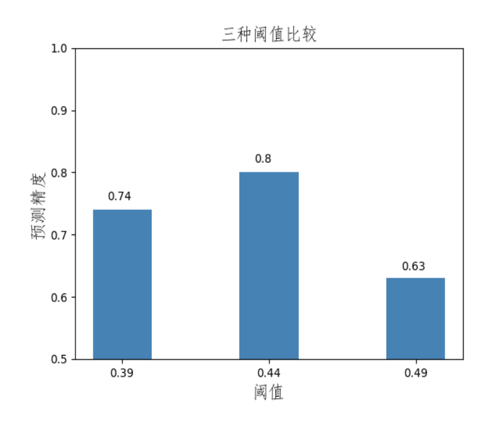


图 7　三种阈值比较情况

令$a_1 - a_7$分别代表道岔故障决策表中的 7 个条件属性，依次为温度，供电电压，发送电压，发送频率，接收 1.1 电压，接收 1.2 电压，干扰电压。如表 5 所示，δ_1和δ_2的约简结果一致，δ_4和δ_5的约简结果也一致，因此应选择$\delta_1, \delta_3, \delta_4$所对应的约简集进行比较。分别将三种约简结果输入到参数设置相同的神经网络中进行训练，每个约简结果进行 10 次预测，取 10 次预测结果平均值，选出其中预测结果较好的属性约简集。如图 7 所示，在经过多次神经网络预测后，阈值$\delta = 0.44$对应的预测精度最高，所以应选择条件属性为$\{a_1, a_2, a_3, a_4\}$(即温度、供电电压、发送电压、发送频率)的约简样本集输入神经网络。

4.4　两种属性约简算法对比

为进一步验证邻域近似条件熵属性约简算法的有效性，下面分别就差别矩阵属性约简算法[错误!未找到引用源。]和邻域近似条件熵属性约简算法从属性约简数目和预测精度两个角度进行对比。其中差别矩阵属性约简是在 equal frequency binning（等频分箱）数据离散化操作的基础上完成的，具体实验结果如表 6~8 所示：

表 6 两种算法约简个数比较

数据集	约简后条件属性个数	
	文献[10]算法	本文算法
故障决策表	3	4

表 7 两种算法约简得到的条件属性

数据集	约简后得到的条件属性	
	文献[10]算法	本文算法
故障决策表	$\{a_1, a_2, a_6\}$	$\{a_1, a_2, a_3, a_4\}$

表 8 BP 神经网络模型下两种约简算法预测准确率对比

数据集	约简后条件属性个数	
	文献[10]算法	本文算法
故障决策表	74%	92%

通过上述实验结果可知邻域近似条件熵属性约简算法能有效的约简样本集中的冗余属性，同时模型的预测精度也远高于差别矩阵属性约简算法。因此邻域近似条件熵属性约简算法是

一种更加有效的约简方法。

4.5 BP 神经网络预测模型

随着深度学习研究的不断深入，很多学者针对特定问题设计出针对性的深度神经网络，主要包括 CNN（卷积神经网络）、RNN（循环神经网络）、LSTM（长短期记忆网络）以及近期研究热点——GANs（生成对抗网络）等。其中 CNN 适用于处理图像数据，RNN 常用于解决语言识别和机器翻译问题，LSTM 适合处理和预测时间序列中间隔和延迟相对较长的事件，而 GANs 可进行图文转换以及图片降噪等操作。由于目前铁路电务设备监测指标（属性）较少，通过系统监测获取到的数据规模相对较小并且数据复杂度也相对较低，因此应用上述四种网络进行实施故障预测并不能提升预测模型的预测精度，而且还有可能在模型学习阶段耗费大量的时间。而 BP 神经网络因其自身结构特征，极其适合处理较小规模的预测问题，故本文选用 BP 神经网络搭建故障预测模型。

考虑到实际应用中电务设备对应的样本数据属性较少，样本理想分类函数复杂性较低，故在设计电务设备故障预测模型时优先选择 3 层 BP 神经网络模型。在经过 K 均值样例约简和邻域近似条件熵属性约简后，样例的条件属性集有 4 个属性，因此神经网络的输入层神经元节点数为 4。由于故障预测问题是一个二分类问题，所以输出层神经元节点数为 1。经过不断的训练和对比，最终确定隐含层神经元节点数为 30 时，神经网路的收敛速度和预测精度达到了预期效果。此外，将神经网络隐含层激活函数设置成 ReLu 来加快故障预测模型的训练速度，通过设置 Sigmoid 函数为输出层的激活函数，确保模型输出结果保持在 $[0,1]$ 之间。

当故障预测模型训练结束后，选取 50 组测试样本输入故障预测模型，评估故障预测模型的性能，最终得到 50 组测试集的故障预测结果（如表 6 所示）。由表 6 数据可知，在 50 组测试样本中有 4 个测试样本出现了错分情况，分别是编号为 2、10、41 和 44 的样本。

表 9 测试样本集的预测结果

编号	预测结果	真实类别	编号	预测结果	真实类别
1	0.000002	0	26	0.999999	1
2	0.998718	0	27	0.006536	0
3	0.000186	0	28	0.999983	1
4	0.628283	1	29	0.887141	1
5	0	0	30	0.000178	0
6	1	1	31	0.000034	0
7	0.893361	1	32	0.999993	1
8	0.999974	1	33	0	0
9	0.000072	0	34	1	1
10	0.998752	0	35	0.00021	0
11	0	0	36	0.000244	0
12	0.00018	0	37	0.999561	1
13	1	1	38	0.999565	1
14	0.8841	1	39	0.000091	0
15	0.99957	1	40	0.999569	1
16	1	1	41	0.99856	0
17	1	1	42	0.87962	1
18	1	1	43	0.60502	1
19	0.99956	1	44	0.99874	0
20	1	1	45	0.00642	0
21	1	1	46	0.88566	1
22	1	1	47	0.64724	1
23	1	1	48	1	1
24	0.580633	1	49	0.90369	1
25	0.999566	1	50	1	1

5 结束语

文中针对铁路电务设备故障频发、运行效率低且无有效故障预测方法等现实问题，提出一种基于 K-均值聚类、邻域近似条件熵、以及 BP 神经网络的电务设备故障预测模型，能够为后续电务设备智能管理提供有效思路和方法。

目前在设备故障预测阶段主要使用设备日常监测数据进行故障的预测。但铁路运输调度的不同，不同设备发生故障后对铁路运输造成的损失不同。所以下一步的研究应额外设置设备故障损失决策表，将设备故障引起的损失和设备故障发生概率集成，最终形成设备维修紧急程度列表，电务维修人员可按照紧急程度列表去检修设备，最大程度地降低由于设备故障所造成的人员和财产损失。

同时，随着铁路智能技术的不断革新，未来的铁路智能系统可以对电务设备实施更全方位的监测，能够获取到维度更多、复杂度更高的数据。同时，随着系统的普及应用，获取的数据量也越来越大。本文所提的 BP 神经网络技术对多维复杂大量数据的处理效率会不尽如意，需要借助于深度神经网络学习更多的复杂数据所蕴含的特征，从而实现电务设备故障的高精准预测。

参考文献：

[1] 龚原斌. 浅谈铁路电务系统故障应急处置存在问题及对策 [J]. 铁道通信信号, 2012, 48 (01): 46-47. (Gong Yuanbin. Discussion about the problem and strategy according to emergency fault handling of railway electrical system [J]. Railway Signaling & Communication, 2012, 48 (01): 46-47.)

[2] 艾红, 周东华. 动态系统的故障预测方法 [J]. 华中科技大学学报: 自然科学版, 2009, 37 (S1): 222-225. (Ai hong, Zhou Donghua. Fault prediction approach for dynamic system [J]. Journal of Huazhong University of Science and Technology: Natural Science Edition, 2009, 37 (S1): 222-225.)

[3] 朱大奇. 人工神经网络原理及应用 [M]. 北京: 科学出版社, 2006. (Zhu Daqi. Principle and application of artificial neural network [M]. Beijing:

Science Press, 2006.)

[4] 郭宇，杨育. 基于灰色粗糙集与BP神经网络的设备故障预测 [J]. 计算机应用研究，2017, 34 (09): 2642-2645. (Guo Yu, Yang Yu. Equipment fault prediction based on grey rough set and BP neural network [J]. Application Research of Computes, 2017, 34 (09): 2642-2645.)

[5] 吴夙慧，成颖，郑彦宁，等. K-means 算法研究综述 [J]. 数据分析与知识发现, 2011, 27 (05): 28-35. (Wu Suhui, Cheng Ying, Zheng Yanning, *et al.* Survey on K-means Algorithm [J]. New Technology of Library and Information Service, 2011, 27 (05): 28-35.)

[6] 胡伟. 改进的层次 K 均值聚类算法 [J]. 计算机工程与应用，2013, 49 (02): 157-159. (Hu Wei. Improved hierarchical K-means clustering algorithm [J]. Computer Engineering and Applications, 2013, 49 (02): 157-159.)

[7] 张宁，范年柏. 基于邻域近似条件熵的启发式属性约简 [J//OL]. 计算机应用研究，2018, 5 (35): 1-2. (Zhang ning, Fan Nianbai. Heuristic attribute reduction based on neighborhood approximate conditional entropy [J]. Application Research of Computes, 2018, 5 (35): 1-2.)

[8] 黄国顺. 保正域的决策粗糙集属性约简 [J]. 计算机工程与应用，2016, 52 (2): 165-169. (Huang Guoshun. Positive region preservation reducts in decision-theoretic rough set models [J]. Computer Engineering and Applications, 2016, 52 (2): 165-169.)

[9] 葛浩，李龙澍，杨传健. 差别矩阵约简表示及其快速算法实现 [J]. 控制与决策, 2016, 31 (1): 12-20. (Ge Hao, Li Longshu, Yang Chuanjian. Discernibility matrix-based reduct representation and quick algorithms [J]. Control and Decision, 2016, 31 (1): 12-20.)

[10] 续欣莹，刘海涛，谢珺，等. 信息观下基于不一致邻域矩阵的属性约简 [J]. 控制与决策, 2016, 31 (1): 130-136. (Xu Xinying, Liu Haitao, Xie Jun, *et al.* Attribute reduction based on inconsistent neighborhood matrix under information view [J]. Control and Decision, 2016, 31 (1): 130-136.)

[11] 胡清华，赵辉，于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法 [J]. 模式识别与人工智能，2008, 21 (06): 732-738. (Hu Qinghua, Zhao Hui, Yu Daren. Efficient symbolic and numerical attribute reduction with neighborhood rough sets [J]. Pattern recognition and artificial intelligence, 2008, 21 (06): 732-738.)

[12] 谢玲玲，雷景生，徐菲菲. 基于改进的邻域粗糙集与概率神经网络的水电机组振动故障诊断 [J]. 上海电力学院学报，2016, 32 (2): 181-187. (Xie Lingling, Lei Jingshen, Xu Feifei. Vibrant fault diagnosis for hydro-turbine generating unit based on improved neighborhood rough sets and PNN [J]. Journal of Shanghai University of Electric Power, 2016, 32 (2): 181-187.)

[13] 徐章艳，刘作鹏，杨炳儒，宋威. 一个复杂度为 max (O (|C||U|) , O (|C|~2|U//C|) 的快速属性约简算法 [J]. 计算机学报，2006, 29 (03): 391-399. (Xu Zhangyan, Liu Zuopeng, Yang Bingru, etal. A quick attribute reduction algorithm with complexity of max (O (|C||U|) , O (|C|~2|U//C|) [J]. Chinese Journal of Computers, 2006, 29 (03): 391-399.)